WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

Investigating Smooth Multiple Regression by the Method of Average Derivatives

Wolfgang Härdle
Thomas M. Stoker

WP# ~~2004~~ 2028-88 February 1987, revised April 1988

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

Investigating Smooth Multiple Regression by
the Method of Average Derivatives

Wolfgang Härdle
Thomas M. Stoker

WP# ~~2004~~ 2028-88 February 1987, revised April 1988

INVESTIGATING SMOOTH MULTIPLE REGRESSION BY THE METHOD OF AVERAGE DERIVATIVES[+]

by Wolfgang Härdle[*] and Thomas M. Stoker[**]

February 1987, revised April 1988

*) Institut für Wirtschaftstheorie II
Adenauerallee 24-26
Universität Bonn
D-5300 Bonn 1, West Germany

**) MIT School of Management, E52-455
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA

+) This research was funded by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303, and a grant from the National Science Foundation. The authors wish to thank R. Carroll, D. Scott, J. Powell, R. Plantiko and several seminar audiences for helpful comments.

ABSTRACT

Let $(x_1, \ldots, x_k, y)$ be a random variable where y denotes a response on the vector x of predictor variables. In this paper we propose a technique (termed ADE) for studying the mean response $m(x)=E(y|x)$ through the estimation of the k-vector of average derivatives $\delta=E(m')$. The ADE procedure involves two stages: first estimate δ using an estimator $\hat{\delta}$, and then estimate $m(x)$ as $\hat{m}(x)=\hat{g}(x^T\hat{\delta})$, where $\hat{g}$ is an estimator of the univariate regression of y on $x^T\hat{\delta}$. We argue that the ADE procedure exhibits several attractive characteristics: data summarization through interpretable coefficients, graphical depiction of the possible nonlinearity between y and $x^T\hat{\delta}$, and theoretical properties consistent with dimension reduction. We motivate the ADE procedure using examples of models that take the form $m(x)=\tilde{g}(x^T\beta)$. In this framework, δ is shown to be proportional to β, and $\hat{m}(x)$ infers $m(x)$ exactly.

The focus of the procedure is on the estimator $\hat{\delta}$, which is based on a simple average of kernel smoothers, and is shown to be a $\sqrt{N}$ consistent and asymptotically normal estimator of δ. The estimator $\hat{g}(.)$ is a standard kernel regression estimator, and is shown to have the same properties as the kernel regression of y on $x^T\delta$. In sum, the estimator $\hat{\delta}$ converges to δ at the rate typically available in parametric estimation problems, and $\hat{m}(x)$ converges at the optimal one-dimensional nonparametric rate.

We study the ADE estimators using Monte Carlo analysis, using sample designs with k=4 predictor variables. The ADE estimators perform well in samples of size N=50 generated from a linear models, and samples of size N=100 generated by a highly nonlinear model. For the latter samples, the ADE procedure is seen to have desirable goodness-of-fit and data summarization features relative to a multivariate regression smoother.

INVESTIGATING SMOOTH MULTIPLE REGRESSION BY THE METHOD OF AVERAGE DERIVATIVES

by Wolfgang Härdle and Thomas M. Stoker

1. Introduction

The popularity of linear modeling in empirical analysis is based, in large part, on the ease with which the results can be interpreted. This tradition influenced the modeling of various parametric nonlinear regression relationships, where the mean response variable is assumed to be a nonlinear function of a weighted sum of the predictor variables. As in linear modeling, this feature is attractive because the coefficients, or weights of the sum, give a simple picture of the relative impacts of the individual predictor variables on the response variable. In this paper we propose a flexible method of studying general multivariate regression relationships in line with this approach. Our method is to first estimate a specific set of coefficients, termed average derivatives, and then compute a (univariate) nonparametric regression of the response on the weighted sum of predictor variables.

The central focus of this paper is analysis of the average derivative, which is defined as follows. Let $(x,y)=(x_1,\ldots,x_k,y)$ denote a random variable, where y is the response studied. If the mean response of y given x is denoted as

$$(1.1)\qquad m(x) = E(y|x)$$

then the vector of "average derivatives" is given as

$$(1.2)\qquad \delta = E(m')$$

where $m' \equiv \partial m/\partial x$ is the vector of partial derivatives, and expectation is taken with respect to the marginal distribution of x. We argue in the next section

that δ represents sensible "coefficients" of changes in x on y.

We construct a nonparametric estimator $\hat{\delta}$ of δ, based on an observed random sample (x_i, y_i), $i=1,\ldots,N$. Our procedure for modeling m(x) is to first compute $\hat{\delta}$, form the weighted sum $\hat{z}_i = x_i^T\hat{\delta}$ for $i=1,\ldots,N$ (where x^T is the transpose of x), and then compute the (Nadaraya-Watson) kernel estimator $\hat{g}(.)$ of the regression of y_i on $\hat{z}_i$. The regression function m(x) is then modeled as

$$(1.3) \qquad \hat{m}(x) = \hat{g}(x^T\hat{\delta})$$

The output of the procedure is three-fold: a summary of the relative impacts of changes in x on y (via $\hat{\delta}$), a graphical depiction of the possible nonlinearity between y and the weighted sum $x^T\hat{\delta}$ (a graph of $\hat{g}$), and a formula for computing estimates of the mean response m(x) (from equation (1.3)). We refer to this as the ADE method, for "average derivative estimation."

The exposition is designed to show that the ADE method has three attractive features: data summarization through interpretable coefficients, computational simplicity and theoretical properties consistent with dimension reduction. The statistic $\hat{\delta}$ is based on a simple average of nonparametric kernel smoothers, and its properties depend only on regularity properties on the joint density of (x,y), or in particular, on no functional form assumptions on the regression function m(x). The limiting distribution of $\sqrt{N}(\hat{\delta}-\delta)$ is multivariate normal. The nonparametric regression estimator $\hat{m}(x)=\hat{g}(x^T\hat{\delta})$ is constructed from a k-dimensional predictor variable, but it achieves the optimal rate $N^{2/5}$ that is typical for one-dimensional smoothing problems (see Stone 1980). While $\hat{\delta}$ and $\hat{g}(.)$ each involve choice of a smoothing parameter, they are computed directly from the data in two steps, and thus require no computer intensive iterative techniques for finding optimal objective function values.

Section 2 motivates the ADE method through several applied examples.

Section 3 introduces the estimators $\hat{\delta}$ and $\hat{g}$ and establishes their large sample statistical properties. Section 4 discusses the results, including the relationship of the ADE method to projection pursuit regression (PPR) of Friedman and Stuetzle(1981) and other flexible methods. The Monte Carlo study of Section 5 shows the ADE method to be well behaved in finite samples. Section 6 follows with some concluding remarks.

2. Motivation of the ADE Procedure

The average derivative δ is most naturally interpreted in situations where the influence of x on y is modeled via a weighted sum $x^T\beta$, where β is a vector of coefficients: where the regression function is $m(x)=\tilde{g}(x^T\beta)$. In such a model, there is an intimate relationship between the coefficients β and the average derivative δ. In particular, $m' = [d\tilde{g}/d(x^T\beta)]\ \beta$, so that $\delta = E[d\tilde{g}/d(x^T\beta)]\beta = \gamma\beta$, where γ is a scalar (assumed nonzero). Consequently, when the mean response is a function of a weighted sum $x^T\beta$, the average derivative δ is a constant multiple of coefficient vector β.

An obvious example is the classical linear regression model: $y = \alpha + x^T\beta + e$, where e is a random variable uncorrelated with x, which gives $\delta=\beta$. Another class of models is those that are linear up to transformations:

$$\phi(y) = \psi(x^T\beta) + e \qquad (2.1)$$

where $\psi(.)$ is a non-constant transformation, $\Phi(.)$ is invertible, and e is a random disturbance that is independent of x. Here we have that $m(x)=E[\phi^{-1}(\psi(x^T\beta)+e)|x]=\tilde{g}(x^T\beta)$. The form (2.1) includes the model of Box-Cox(1964), where $\phi(y)=(y^{\lambda_1}-1)/\lambda_1$ and $\psi(x^T\beta)=\alpha+[(x^T\beta)^{\lambda_2}-1]/\lambda_2$.

Other models exhibiting this structure are discrete regression models, where y is 1 or 0 according to

$$
(2.2) \qquad \begin{aligned} y &= 1 && \text{if } e < \Psi(x^T\beta) \\ &= 0 && \text{if } e \geq \Psi(x^T\beta) \quad . \end{aligned}
$$

Here the regression function $m(x)$ is the probability that $y=1$, which is given as $m(x)=\mathrm{Prob}\{e<\Psi(x^T\beta)|x\}=\tilde{g}(x^T\beta)$. References to specific examples of binary response models can be found in Manski and McFadden(1981). Standard probit models are included by assuming that $\Psi(x^T\beta)=\alpha+x^T\beta$ and e is a normal random variable (with distribution function Φ), giving $m(x)=\Phi(\alpha+x^T\beta)$. Logistic regression models are likewise included; here $m(x)=\exp(\alpha+x^T\beta)/[1+\exp(\alpha+x^T\beta)]$.

Censored regression, where

$$
(2.3) \qquad \begin{aligned} y &= \Psi(x^T\beta) + e && \text{if } \Psi(x^T\beta) + e \geq 0 \\ &= 0 && \text{if } \Psi(x^T\beta) + e < 0 \end{aligned}
$$

is likewise included, and setting $\Psi(x^T\beta)=\alpha+x^T\beta$ gives the familiar censored linear regression model (see Powell(1986) among others). For further examples from econometric modeling, see Stoker(1986).

A parametric approach to the estimation of any of these models, for instance based on maximum likelihood, requires the (parametric) specification of the distribution of the random variable e and of the transformations $\Psi(.)$, and for (2.1), the transformation $\varphi(.)$. Substantial bias can result if any of these features is incorrectly specified. Nonparametric estimation of $\delta=\gamma\beta$ avoids such restrictive specifications. In fact, the form $m(x)=\tilde{g}(x^T\beta)$ generalizes the "generalized linear models (GLIM)"; see McCullagh and Nelder(1983). These models have $\tilde{g}$ invertible with $\tilde{g}^{-1}$ referred to as the "link" function. Other approaches that generalize GLIM models can be found in Breiman and Friedman(1985) and Hastie and Tibshirani (1986).

Turning our attention to ADE regression modeling, we show in the next section that $\hat{m}(x)$ of (1.3) will estimate $g(x^T\delta)=E(y|x^T\delta)$, in general.

Consequently, the ADE method will completely infer m(x) when

$$m(x) = g(x^T\delta) \quad (2.4)$$

But it is easy to see that (2.4) will always be valid when m(x) takes the form $m(x)=\tilde{g}(x^T\beta)$, as in the above examples. Suppose that such a model were reparameterized to have coefficients $\beta^*=c\beta$, where c is a nonzero scalar, then we can write $m(x)=\tilde{g}^*(x^T\beta^*)$, with $\tilde{g}^*$ defined as $\tilde{g}^*(.)\equiv\tilde{g}(./c)$. Thus when $m(x)=\tilde{g}(x^T\beta)$, we can equivalently reparameterize m(x) to have coefficients $\delta=\gamma\beta$, giving (2.4) with $g(.)=\tilde{g}(./\gamma)$. This corresponds to a normalization exhibited by the function g(.) of (2.4), namely $E[dg/d(x^T\delta)]=1$. To understand this normalization, consider a reparameterization of β to $\beta^*=c\beta$, and note that $\delta=\gamma\beta=\gamma^*\beta^*$, where $\gamma^* = \gamma/c = E[d\tilde{g}^*/d(x^T\beta^*)]$, with $\tilde{g}^*$ defined as above. Setting $c=\gamma$ gives $\beta^*=\gamma\beta=\delta$, and we conclude that $E[dg/d(x^T\delta)] = \gamma^* = \gamma/\gamma = 1$.

Another way of interpreting the scale of δ is to consider the change in the mean of y when x is translated to $x+\Delta x$ in the transformation model (2.1). The average change is proportional to $(\Delta x)^T\delta$, as it is in the linear model. Other scalings of the coefficients β make this change dependent on $\Phi(.)$ and $\Psi(.)$.

3. Kernel Estimation of Average Derivatives

Our approach to estimation of δ utilizes nonparametric estimation of the marginal density of x. Let f(x) denote this marginal density, $f'\equiv\partial f/\partial x$ the vector of partial derivatives, and $\ell \equiv -\partial \ln f/\partial x = -f'/f$ the negative log-density derivative. If f(x)=0 on the boundary of x values, then integration by parts gives

$$\delta = E(m') = E(\ell y) \quad (3.1)$$

Our estimator of δ is a sample analogue of the last term in this formula.

using a nonparametric estimator of $\ell(x)$ evaluated at each observed value x_i, $i=1,\ldots,N$..

In particular, the density function $f(x)$ is estimated at x_i using the (Rosenblatt-Parzen) kernel density estimator:

$$(3.2) \qquad \hat{f}_h(x) = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{1}{h}\right]^k K\left[\frac{x - x_j}{h}\right]$$

where $K(.)$ is a kernel function, $h=h_N$ is the bandwidth parameter, and $h \to 0$ as $N \to \infty$. The vector function $\ell(x)$ is then estimated at x_i using $\hat{f}_h(x)$ as

$$(3.3) \qquad \hat{\ell}_h(x) = - \hat{f}_h'(x) / \hat{f}_h(x)$$

where $\hat{f}_h' \equiv \partial \hat{f}_h / \partial x$ is an estimator of the partial density derivative. For a suitable kernel $K(.)$, under general conditions $\hat{f}_h(x)$, $\hat{f}_h'(x)$ and $\hat{\ell}_h(x)$ are consistent estimators of $f(x)$, $f'(x)$ and $\ell(x)$, respectively.

Because of division by $\hat{f}_h(x)$, the function $\hat{\ell}_h(x)$ may exhibit erratic behavior when the value of $\hat{f}_h$ is very small. Consequently, for estimation of δ we only include terms for which the value of $\hat{f}_h(x_i)$ is above a bound. Toward this end, define the indicator $\hat{I}_i = I[\hat{f}_h(x_i) > b]$, where $I[.]$ is the indicator function, and where $b=b_N$ is a trimming bound such that $b \to 0$ as $N \to \infty$.

The "average derivative estimator" $\hat{\delta}$ is defined as:

$$(3.4) \qquad \hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_h(x_i) y_i \hat{I}_i \quad .$$

We derive the large sample statistical properties of $\hat{\delta}$ on the basis of smoothness conditions on $m(x)$ and $f(x)$. The required assumptions (given formally in the Appendix) are described as follows. As above, the k-vector x is continuously distributed with density $f(x)$, and $f(x)=0$ on the boundary of x values. The regression function $m(x)=E(y|x)$ is (a.e.) continuously differentiable, and the second moments of m' and ℓy exist. The density $f(x)$ is

assumed to be smooth, having partial derivatives of order $p \geq k+2$. The kernel function $K(.)$ has compact support and is assumed to be of order p. Finally, we assume some technical conditions on the behavior of $m(x)$ and $f(x)$ in the tails of the distribution: for instance ruling out thick tails, and rapid increases in $m(x)$ as $|x| \to \infty$.

Under these conditions, $\hat{\delta}$ is an asymptotically normal estimator of δ, stated formally as

Theorem 3.1: Given Assumptions 1.-9. stated in the Appendix, if

(i) $N \to \infty$, $h \to 0$, $b \to 0$, $b^{-1}h \to 0$.

(ii) For some $\varepsilon > 0$, $b^4 N^{1-\varepsilon} h^{2k+2} \to \infty$.

(iii) $Nh^{2p-2} \to 0$.

then $\sqrt{N}(\hat{\delta} - \delta)$ has a limiting normal distribution with mean 0 and variance Σ, where Σ is the covariance matrix of $r(y,x)$, with

$$(3.5) \qquad r(y,x) = m'(x) + [y - m(x)]\ell(x)$$

The proof of Theorem 3.1, as well as those of the other results of the paper, are contained in the Appendix.

For the purpose of carrying out inference on the value of δ, the covariance matrix Σ could be consistently estimated as the sample variance of uniformly consistent estimators of $r(y_i,x_i)$, $i=1,\ldots,N$, and the latter could be constructed using any uniformly consistent estimators of $\ell(x)$, $m(x)$ and $m'(x)$. The proof of Theorem 3.1 suggests a more direct estimator of $r(y_i,x_i)$, defined as

(3.6)

$$\hat{r}_{hi} = \hat{\ell}_h(x_i)y_i\hat{I}_i + \frac{1}{N}\sum_{j=1}^{N}\left[h^{-k-1}K'\left[\frac{x_i-x_j}{h}\right] - h^{-k}K\left[\frac{x_i-x_j}{h}\right]\hat{\ell}_h(x_j)\right]\frac{y_j\hat{I}_j}{\hat{f}_h(x_j)}$$

Define the estimator $\hat{\Sigma}$ of Σ as

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\hat{r}_{hi}\hat{r}_{hi}^T\hat{I}_i - \hat{\delta}\hat{\delta}^T \tag{3.7}$$

We then have

Theorem 3.2: If $N\to\infty$, $h\to 0$, $b\to 0$ and $b^{-1}h\to 0$, then $\hat{\Sigma}$ is a consistent estimator of Σ.

Theorem 3.2 facilitates inference on hypotheses about δ. For instance, consider testing restrictions that certain components of δ are zero, or equality restrictions across components of δ. Such restrictions are captured by the null hypothesis that $Q\delta=\delta_0$, where Q is a $k_1\times k$ matrix of full rank $k_1\le k$. Tests of this hypothesis can be based on the Wald statistic $W=N(Q\hat{\delta}-\delta_0)^T(Q\hat{\Sigma}Q^T)^{-1}(Q\hat{\delta}-\delta_0)$, which has a limiting χ^2 distribution with k_1 degrees of freedom.

We now turn our attention to the estimation of $g(x^T\delta)=E(y|x^T\delta)$, and add the assumption that $g(.)$ is twice differentiable. Set $\hat{z}_j=x_j^T\hat{\delta}$, $j=1,\ldots,N$, and let f_1 denote the density of $z=x^T\delta$. Define $\hat{g}(z)$ as the (Nadaraya-Watson) kernel estimator of the regression of y on $\hat{z}=x^T\hat{\delta}$:

$$\hat{g}_{h'}(z) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{h'}\right]K_1\left[\frac{z-\hat{z}_j}{h'}\right]y_j \Bigg/ \hat{f}_{1h'}(z) \tag{3.8}$$

where $\hat{f}_{1h'}$ is the density estimator:

$$\hat{f}_{1h'}(z) = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{1}{h'}\right] K_1\left[\frac{z - \hat{z}_j}{h'}\right] \tag{3.9}$$

with bandwidth $h'=h'_N$, and where K_1 is a symmetric (positive univariate) kernel function. Suppose, for a moment, that $z_j=x_j^T\delta$ instead of $\hat{z}_j$ were used in the (3.8) and (3.9), then it is well known (Schuster(1972)) that the resulting regression estimator is asymptotically normal and converges (pointwise) at the optimal (univariate) rate $N^{2/5}$. Theorem 3.3 states that there is no cost to using the estimated values $\hat{z}_j$ as above:

<u>Theorem 3.3</u>: Suppose z is such that $f_1(z)\geq b_1>0$. If $N\to\infty$, $h'\sim N^{-1/5}$, then $N^{2/5}[\hat{g}_{h'}(z)-g(z)]$ has a limiting normal distribution with mean B(z) and variance V(z), where

$$\begin{aligned} B(z) &= \frac{1}{2} [g''(z) + 2 g'(z)f_1'(z)/f_1(z)] \int u^2K_1(u)\, du \\ V(z) &= [\mathrm{Var}(y|x'\delta=z)/f_1(x)] \int K_1(u)^2\, du \end{aligned} \tag{3.10}$$

The bias and variance given in (3.10) can be estimated consistently for each z using y, $\hat{g}_{h'}$ and $\hat{f}^z_{h'}$ and their derivatives, using standard methods. Therefore, asymptotic confidence intervals can be constructed for $\hat{g}_{h'}(z)$. It is clear that the same confidence intervals apply to $\hat{m}(x)=\hat{g}_{h'}(x^T\hat{\delta})$, for $z=x^T\hat{\delta}$.

4. Remarks and Discussion

4.1 On The Average Derivative Estimator

As indicated in the introduction, the most interesting feature of Theorem 3.1 is that $\hat{\delta}$ converges to δ at rate $\sqrt{N}$. This is the rate typically available in parametric estimation problems, and is the rate that would be attained if the values $\ell(x_i)$, i=1,...,N were known and used in the average (3.4). The

estimator $\hat{\ell}_h(x)$ converges pointwise to $\ell(x)$ at a slower rate, so Theorem 3.1 gives a situation where the average of nonparametric estimators converges more quickly than any of its individual components. This occurs because of the overlap between kernel densities at different evaluation points; for instance, if x_i and x_j are sufficiently close, the data used in the local average $\hat{f}_h(x_i)$ will overlap with that used in $\hat{f}_h(x_j)$. These overlaps lead to the approximation of $\hat{\delta}$ by U-statistics with kernels depending on N. The asymptotic normality of $\hat{\delta}$ follows from results on the equivalence of such U-statistics to (ordinary) sample averages. In a similar spirit, Powell, Stock and Stoker(1987) obtain $\sqrt{N}$ convergence rates for the estimation of "density weighted" average derivatives, and Robinson(1987) and Härdle and Marron(1987) show how kernel densities can be used to obtain $\sqrt{N}$ convergence rates for certain parameters in specific semiparametric models. We also note that our method of trimming follows Bickel(1982), Manski(1984) and Robinson(1986).

For any given sample size, the bandwidth h and the trimming bound b can be set to any (positive) values, so that their choice can be based entirely on the small sample behavior of $\hat{\delta}$. The conditions (i)-(iii) of Theorem 3.1 indicate how the initial bandwidth and trimming bound must be decreased as the sample size is increased. These conditions are certainly feasible: suppose $h=h_0N^{-\zeta}$ and $b=b_0N^{-\eta}$, then (i)-(iii) are equivalent to

$$\zeta > \eta > 0$$

$$\frac{1}{2p-2} < \zeta < \frac{1-4\eta-\varepsilon}{2k+2}$$

Since $p \geq k+2$ and ε is arbitrarily small, η can be chosen small enough to fulfil the last condition.

The bandwidth conditions arise as follows. Condition (ii) assures that the estimator $\hat{\delta}$ can be "linearized" to one without an estimated denominator, and is a sufficient condition for asymptotic normality. Condition (iii)

assures that the bias of $\hat{\delta}$ vanishes at rate $\sqrt{N}$. Conditions (i)-(iii) are one-sided in implying that the trimming bound b cannot converge too quickly to 0 as $N\to\infty$, but rather must converge slowly. The behavior of the bandwidth h as $N\to\infty$ is bounded both below and above by conditions (ii) and (iii).

Condition (iii) does imply that the pointwise convergence of $\hat{f}_h(x)$ to f(x) must be suboptimal. Stone(1980) shows that the optimal pointwise rate of convergence under our conditions is $N^{p/(2p+k)}$, and Collomb and Härdle(1986) show that this rate is achievable with kernel density estimators such as (3.2): for instance, by taking $h_{opt}=h_0N^{-1/(2p+k)}$. But we have that $Nh_{opt}^{2p-2}\to\infty$, which violates condition (iii), so that as $N\to\infty$, h must converge to 0 more quickly than h_{opt}. The reason for this is that (iii) is a bias condition: as $N\to\infty$, the (pointwise) bias of $\hat{f}_h(x)$ must vanish at a faster rate than its (pointwise) variance, for the bias of $\hat{\delta}$ to be $o(N^{-1/2})$. In other words, for $\sqrt{N}$ consistent estimation of δ, one must "undersmooth" the nonparametric component $\hat{\ell}_h(x)$.

4.2 On Modeling Multiple Regression

The main implication of Theorem 3.3 is that the optimal one-dimensional convergence rate is achievable in the estimation of $g(x^T\delta)=E(y|x^T\delta)$, using $\hat{\delta}$ instead of δ. We have assumed that m(x), and hence $g(x^T\delta)$, is twice differentiable, but this plays no role except to affix the optimal rate at $N^{2/5}$. If $g(x^T\delta)$ is assumed to be differentiable of order q and $K_1()$ is a kernel of order q, then the result is valid, where the optimal rate of convergence is then $N^{q/(2q+1)}$. The attainment of optimal one-dimensional rates of convergence is possible for the ADE method because the additive structure of $g(x^T\delta)$ is sufficient to follow the "dimension reduction principle" of Stone(1986). Alternative uses of additive structure can be found in Breiman and Friedman (1985) and Hastie and Tibshirani (1986).

Finally, we turn to the connection between the ADE method and projection pursuit regression (PPR) of Friedman and Stuetzle(1981). The first step of PPR is to choose β (normalized as a direction) and $\tilde{g}$ to minimize $s(\tilde{g},\beta)=\Sigma\,[y_i - \tilde{g}(x_i^T\beta)]^2$. Because any model of the form $m(x)=\tilde{g}(x^T\beta)$ is fully inferred by $\hat{m}(x)=\hat{g}_{h'}(x^T\hat{\delta})$ at the optimal one dimensional rate of convergence, one can correctly regard the ADE method as a version of projection pursuit regression. However, for a general regression function $m(x)$, $\hat{g}$ and $\hat{\delta}$ will not necessarily minimize the sum of squares $s(\tilde{g},\beta)$. The first order conditions of that minimization implies that given $\tilde{g}$, β is chosen such that $\{y_i-\tilde{g}(x_i^T\beta)\}$ is orthogonal to $\{x_i\tilde{g}'(x_i^T\beta)\}$, which does not necessarily yield $\beta=\hat{\delta}/|\hat{\delta}|$.

Given $\hat{\delta}$, the ADE method utilizes a local least squares estimator $\hat{g}_{h'}$; in particular $\sum K^z[(z-x_i^T\hat{\delta})/h'](y_i-t)^2$ is minimized by $t=\hat{g}_{h'}(z)$. Moreover, $\hat{\delta}$ can be seen to be a type of least squares estimator, as follows. Let S^ℓ denote the sample moment $S_\ell=N^{-1}\Sigma\hat{\ell}_{hi}(x_i)\hat{\ell}_{hi}(x_i)^T\hat{I}_i$, and set $\hat{L}_i=(S_\ell)^{-1}\hat{\ell}_{hi}(x_i)\hat{I}_i$. Then $\hat{\delta}$ can easily be seen to be the value of d that minimizes the sum-of-squares $\Sigma[y_i - \hat{L}_i^T d]^2$. Thus $\hat{\delta}$ is chosen such that $\{y_i - \hat{L}_i^T\hat{\delta}\}$ is orthogonal to the subspace spanned by $\{\hat{L}_i\}$, or equivalently, $(S_\ell)^{-1}\hat{\delta}$ represents the coordinates of $\{y_i\}$ projected onto the subspace spanned by $\{\hat{\ell}_{hi}(x_i)\hat{I}_i\}$.

Consequently, when the true regression function is of the form $m(x)=\tilde{g}(x^T\beta)$, ADE and PPR represent different methods of inferring $m(x)$. The possible advantages of the ADE method arise from reduced computational effort, since (given h, b and h') $\hat{m}(x)=\hat{g}(x^T\hat{\delta})$ is computable directly from the data. The first step of PPR will in principle estimate $m(x)$, but minimizing $s(\tilde{g},\beta)$ by an iterative numerical process (of checking all potential directions β and determining the optimal $\tilde{g}$ for each β) typically involves considerable computational effort (although the results of Ichimura (1987) may provide some improvement).

5. The Average Derivative Method in Practice

In this section we present the results of a simulation study designed to study three issues of the ADE approach:

1. The performance of the estimators when the data is generated by a simple parametric model.
2. The ability of the ADE approach to nonparametrically capture structures in high dimensions.
3. The value of dimension reduction, as expressed through the one-dimensional rate of convergence of $\hat{m}(x)$.

The first issue is studied by generating data from a true linear regression model and estimating the coefficients by the ADE method. The second issue is studied using data generated by a highly nonlinear regression model (a sine wave). The third issue is addressed by comparing the final ADE regression estimator with a multivariate nonparametric regression smoother. The simulations utilize k=4 predictor variables and sample sizes of N=50 and N=100.

Our Monte Carlo experience indicated that for relatively small sample sizes, better performance of $\hat{\delta}$ was obtained using a standard positive kernel instead of the higher order kernel prescribed by Theorem 3.1 (for k=4, a kernel of order p=6 is indicated). In particular, the order of the kernel affects the tradeoff between bias and variance, with the higher order kernel required for the bias to vanish at rate $\sqrt{N}$. However, for sample sizes in the range of N=100, we found the small sample bias-variance tradeoff quite pronounced: using a standard positive kernel instead of a higher order kernel implied a slightly higher bias in $\hat{\delta}$ but a substantially smaller variance in most cases.

For the estimator $\hat{\delta}$, the kernel function K was constructed as the product of one dimensional kernels, namely $K(u_1,\ldots,u_k)=\Pi_{j=1}^{k}K_1(u_j)$, where K_1 is taken to be the univariate "biweight" kernel;

$$K_1(u) = (15/16)\ (1 - u^2)^2\ I(|u| \le 1)$$

For estimating $\hat{g}$, we also utilized the biweight kernel.

Our theoretical results do not constrain the choice of bandwidth h or trimming bound b for a given sample size, and we study the behavior of $\hat{\delta}$ for different bandwidth values. We did find that somewhat better estimator behavior resulted when a slightly higher bandwidth value was used to estimate density derivatives than the density value, and so results on $\hat{\delta}$ for bandwidth h have used density derivative estimates computed using $h^*=1.25h$ (so that $\hat{\ell}_h=\hat{f}_{h*}'/\hat{f}_h$). This does not affect the asymptotic approximation theory of Section 3, and is suggested by the slower pointwise convergence rates applicable to the estimation of derivatives (relative to estimation of density levels). Finally, because the value of the trimming bound b is not easily interpreted, we adopted a trimming rule suggested by Ray Carroll, to set b so that a given percentage α of the data was dropped in each sample (we utilize the values α=5% and α=1% below, although the results were not particularly sensitive to the value of α).

5.1 Average Derivative Estimation with a Simple Parametric Model

It is natural to ask how a method that accounts for nonlinearity behaves when the sample is generated by a simple parametric model. In order to study this question we generated samples of size N=50 from the linear model

$$(5.1) \qquad y_i = \sum_{j=1}^{4} j\, x_{ji} + (.05)e_i$$

so that the coefficient of x_{ji} is j, where $x_{1i},\ldots,x_{4i}$, and e_i are independent

standard normal variables. Here the true average derivative is $\delta=(1,2,3,4)$.

Table 1 gives the means and standard deviations of $\hat{\delta}$ over 100 Monte Carlo samples. The best behavior for $\hat{\delta}$ was found using the bandwidth h=1.5, and we include results using a low bandwidth h=1.0 and a high bandwidth h=2.0 for comparison. We also include the means of the trimming bound b over the samples. While there is not a large amount of "noise" in this design (the standard deviation of the additive disturbance is .05), we find the performance of $\hat{\delta}$ to be good, using 50 observations to estimate k=4 coefficients.

5.2 Average Derivative Estimation with Nonlinear Structure

In order to study the behavior of the ADE estimators in a reasonably challenging setting, we first generated samples of size N=100 according to the model

$$y_i = \sin\left(\sum_{j=1}^{4} x_{ji}\right) + e_i \tag{5.2}$$

where $x_{1i}, \ldots, x_{4i}$ and e_i are standard normal variables as before. For this model we have $m'=\cos(\Sigma_j x_{ji})$, so that $\delta=E(m')$ takes the form $\delta=\delta_0(1,\ldots,1)$ with $\delta_0=.135$ (using formulae given in Bronstein-Semandjajew(1974, p.350)).

Table 2 gives the means and standard deviations of the components of $\hat{\delta}$ over 100 Monte Carlo samples. Here we found the coefficients to be well-estimated using bandwidth h=.9, and again we include low and high bandwidth results for comparison. Finally, we present the means and standard deviations of the "estimator" δ^+ that uses the true log-density derivative $\ell(x)$ instead of the estimator $\hat{\ell}_h(x)$ (namely $\delta^+=N^{-1}\Sigma y_i\ell(x_i)$, where $\ell(x_i)=x_i$ for x distributed normally with mean 0 and identity variance-covariance matrix).

It is clear that for the bandwidth h=.9 the estimator $\hat{\delta}$ performs quite well, exhibiting comparable behavior to the "known density" statistic δ^+. The

substantial variation of δ^+ is indicative of the high level of noise in the basic design (the standard error of 1 for e). At any rate, we find the behavior of $\hat{\delta}$ to be encouraging, given k=4 predictor variables and N=100 data points.

Although not reported, we also performed simulations for sample size N=400. We found good behavior of $\hat{\delta}$ for somewhat smaller bandwidth values (in the range of h=.6), as well as diminished standard errors consistent with the quadrupling of the sample size.

5.3 Dimension Reduction of Average Derivative Estimation

The main consequence of Theorem 3.3 is that the ADE regression estimator exhibits one-dimensional rates of convergence. We study this effect by comparing the ADE regression estimator $\hat{m}(x)=\hat{g}_{h'}(x^T\hat{\delta})$ to a high dimensional smoother, namely the Nadaraya-Watson kernel regression estimator $\tilde{m}_{h''}(x)$. We utilize the data sets of size N=100 from the sine model (5.2). We utilize the bandwidths h=.9 for computing $\hat{\delta}$ (and α=5%), and h"=1.3 for the multivariate smoother $\tilde{m}_{h''}$. For computing $\hat{g}_{h'}$, we utilize normalized coefficients $\hat{\delta}^*=\hat{\delta}/|\hat{\delta}|$, and bandwidth h'=.3. The values of h" and h' were determined by cross validation applied to the first few Monte Carlo samples (c.f. Härdle and Marron (1985)).

We begin by studying the overall fit of the ADE estimator versus the multivariate smoother. For this, we compute the difference in the average squared error defined as

$$DIF = ASE_{\tilde{m}_{h''}} - ASE_{\hat{m}}$$

where

$$ASE_{\tilde{m}_{h''}} = N^{-1}\sum_{i=1}^{N}[m(x_i) - \tilde{m}_{h''}(x_i)]^2 \quad ; \quad ASE_{\hat{m}} = N^{-1}\sum_{i=1}^{N}[m(x_i) - \hat{g}_{h'}(x_i^T\hat{\delta})]^2$$

Figure 1 shows the variable DIF over 50 Monte Carlo samples. It is apparent that ADE estimator displays smaller average squared error in the majority of cases, with DIF exceeding .1 for 25% of the samples, and less than -.1 for only 10% of the samples.

In addition to the gains in statistical fit from dimension reduction, we have argued that the ADE method is attractive because it permits a graphical depiction (a graph of $\hat{g}_{h'}$) of the nonlinearity exhibited by the data. Figures 2 through 5 illustrate this feature of the approach, as applied to a single Monte Carlo sample from model (5.2).

Figure 2 shows the basic data, by plotting y_i against the true weighted sum $z_i=x_i^T\delta^*$, where $\delta^* = .5(1,1,1,1)$, as well as a plot of the ADE regression estimator $\hat{g}_{h'}$ of the regression of y_i on the $\hat{z}_i=x_i^T\hat{\delta}^*$, where for this example, $\hat{\delta}^*=(.54,.47,.59,.35)$. While the data shows considerable noise, the basic sine wave pattern is evidenced by both the data and the ADE regression estimator.

Figure 3 displays the plot of the ADE regression estimator $\hat{g}_{h'}$, as well as the plot of the true regression $m(x_i)$ against z_i. Figure 4 displays $\hat{g}_{h'}$ as well as the plot of the kernel smoother $\tilde{g}_{h'}$ of the regression of y_i on the true z_i values (using h'=.3). These two figures point out that the ADE regression estimator tracks the true regression reasonably well, and tracks the smoother based on knowledge of the true δ quite well.

Finally, Figure 5 contains a plot of the multivariate smoother $\tilde{m}_{h''}(x_i)$ against z_i, and the true regression $m(x_i)$ against z_i. While one can discern a sine pattern from the multivariate smoother, it yields a considerably less smooth graphical depiction of the relationship. Moreover, the jagged feature of the plot is an artifact of plotting $\tilde{m}_{h''}(x_i)$ against $z_i=x_i^T\delta^*$, as in fact, the estimated function $\tilde{m}_{h''}$ is quite smooth. We experimented with several

values of h", with larger values of h" moving the jagged plot toward the horizontal, and smaller values of h" increasing the heights of the peaks, each obliterating the visual tendency to a sinusoidal shape. In this sense, we regard Figure 5 as giving the best showing for the multivariate smoother technique.

6. Concluding Remarks

In this paper we have argued that the ADE method represents a useful yet flexible tool for studying general regression relationships. At its center is the estimation of average derivatives, which we propose as sensible coefficients for measuring the relative impacts of separate predictor variables on the mean response. As such, we propose the ADE method as a natural outgrowth of linear modeling, or "running (OLS) regressions," as a useful method of data summarization.

As a procedure for inferring regression, the ADE method displays one-dimensional rates of convergence, and as such achieves theoretical dimension reduction. But of greater practical importance is the economy achieved with respect to data summarization. Instead of facing the nontrivial task of interpreting a fully flexible nonparametric regression, the ADE method permits the "significance" of individual predictor variables to be judged via simple hypothesis tests on the value of the average derivatives. Moreover, remaining nonlinearity can be assessed by looking at a graphical picture of the function $\hat{g}$.

While our simulations of the ADE approach are encouraging, there are many future research topics suggested by our results. For instance, there are numerous practical questions concerning choice of bandwidth for the average derivative estimator, such as how to set the bandwidth optimally. While we have seen some basic tendencies in the Monte Carlo simulations, practical

problems are, of course, never so stylized. The same remarks apply to the rules for trimming, although in our Monte Carlo experience we found no substantial effect of varying the trimming rule.

The ADE estimators are implimented as part of the exploratory data software package XploRe of Härdle(1987). In addition, David Scott of Rice University has written a very efficient code (for the S package) for computing the ADE coefficients, which is available from the authors.

Appendix: Assumptions and Proofs of Theorems

Assumptions for Theorem 3.1.

1. The support Ω of f is a convex, possibly unbounded subset of R^k with nonempty interior. The underlying measure of (y,x) can be written as $\nu_y \times \nu_x$, where ν_x is Lebesgue measure.

2. $f(x)=0$ for all $x \in d\Omega$, where $d\Omega$ is the boundary of Ω.

3. $m(x)=E(y|x)$ is continuously differentiable on $\bar{\Omega} \subset \Omega$, where $\bar{\Omega}$ differs from Ω by a set of measure 0.

4. The moments $E(\ell^T \ell y^2)$ and $E[(m')^T(m')]$ exist.

5. All derivatives of f(x) of order p exist, where $p \geq k+2$.

6. The kernel function K(u) has bounded support $S=\{u| \ |u| \leq 1\}$, is symmetric, $K(u)=0$ for $u \in dS=\{u| \ |u|=1\}$, and is of order p:

$$\int K(u)du = 1$$

$$\int u_1^{\ell_1} \cdots u_k^{\ell_{\rho'}} K(u)du = 0 \qquad \ell_1+\ldots+\ell_{\rho'} < p$$

$$\int u_1^{\ell_1} \cdots u_k^{\ell_{\rho'}} K(u)du \neq 0 \qquad \ell_1+\ldots+\ell_{\rho'} = p$$

(Such kernels are easily constructed by multiplication of one dimensional kernels - see Gasser, Müller and Mammitzsch(1985) for examples of such one dimensional kernel functions).

7. The functions f(x) and m(x) obey the following Lipschitz conditions: for ν in an open neighborhood of 0, there exists functions ω_f, $\omega_{f'}$, $\omega_{m'}$ and $\omega_{\ell m}$ such that

$$|f(x+\nu) - f(x)| < \omega_f(x)|\nu|$$

$$|f'(x+\nu) - f'(x)| < \omega_{f'}(x)|\nu|$$

$$|m'(x+\nu) - m'(x)| < \omega_{m'}(x)|\nu|$$

$$|\ell(x+\nu)m(x+\nu) - \ell(x)m(x)| < \omega_{\ell m}(x)|\nu|$$

with $E[(\ell y\omega_f)^2]<\infty$, $E[(y\omega_{f'})^2]<\infty$, $E[\omega_{m'}^2]<\infty$ and $E[\omega_{\ell m}^2]<\infty$. $M_2(x)=E(y^2|x)$ is continuous in x.

Let $A_N=\{x|f(x)>b\}$ and $B_N=\Omega\backslash A_N$.

8. As $N\to\infty$,

$$\int_{B_N} m(x)f'(x)dx = o(N^{-1/2})$$

9. If $f_\iota^{(p)}$ denotes any p^{th} order partial derivative of f, then $f_\iota^{(p)}$ is locally Hölder continuous: there exists $\tilde{c}(x)$ and $\gamma>0$ such that $|f_\iota^{(p)}(x+\nu)-f_\iota^{(p)}(x)|\le\tilde{c}(x)|\nu|^\gamma$. The $p+\gamma$ moments of K(.) exist. Moreover

$$\int_{A_N} m(x)f_\iota^{(p)}(x)dx \le M < \infty$$

$$h^\gamma \int_{A_N} \tilde{c}(x)m(x)dx \le M < \infty$$

$$h\int_{A_N} m(x)\ell(x)f_\iota^{(p)}(x)dx \le M < \infty$$

$$h^{\gamma+1} \int_{A_N} \tilde{c}(x)m(x)\ell(x)dx \le M < \infty$$

Assumptions 8 and 9 are conditions on the behavior of m(x) and f(x) in the tails of the distribution. For Assumption 8, some sufficient conditions for the univariate case k=1 are as follows, with multivariate analogues easy to derive.

The support Ω is the real line R. Let $L_N=\inf\{|x| \mid f(x)\le b\}$, so that $L_N\to\infty$. Consider the situation where $|m(x)f'(x)|\le C\, G(x)$, where G(x) is a density. Then

$$\int_{B_N} m(x)f'(x)dx \le C \int_{|x|>L_N} G(x)dx$$
$$\le C\ \mathrm{Prob}_G\{|x|>L_N\}$$
$$\le C\ E_G[|x|^d]/L_N^d$$

Therefore, if G(x) has absolute moments of order d, where d is such that $\sqrt{N}L_N^{-d} \to 0$, then Assumption 8 obtains. A sufficient condition is that G(x) is a normal density.

<u>The support Ω is bounded</u>. The density f(x) must be smooth and flat near the boundary. Suppose that $\Omega=[0,x_0]$, $f(0)=0$, $f(L_N)=b_N$, f(x) is increasing over $[0,L_N]$, and m(x) is bounded over $[0,L_N]$. If we expand f'(x) in a Taylor series about 0, and assume that $f'(0)=f^{(2)}(0)= \ldots =f^{(q-1)}(0)=0$, then

$$\int_{[0,L_N]} f'(x)dx = O(L_N^q)$$

Consequently, if q is large enough such that $\sqrt{N}L_N^q \to 0$, then Assumption 8 is valid.

Assumption 9 is written in a weak form required for Theorem 3.1. For instance, Assumption 9 is clearly valid if the support Ω is bounded and m(x) is bounded.

To prove Theorem 3.3, we also require that m(x)=E(y|x) is twice differentiable for all x in the interior of Ω.

Proofs of the Main Results

We begin with two preliminary remarks. First, equation (3.1) follows from integration by parts as

$$\delta = \int m'(x)f(x)dx = - \int m(x)f'(x)dx = \int m(x)\ell(x)f(x)dx = E(\ell y)$$

where the boundary terms vanish by Assumption 2 (c.f. Beran(1977), Stoker(1986)) and the last equality is by iterated expectation.

Second, because of condition (iii), as $N\to\infty$, the pointwise mean square errors of $\hat{f}_h$ and $\hat{f}_h'$ are dominated by their variances (as discussed in Section 4.1). Therefore, since the set $\{x|f(x)\geq b\}$ is compact and $b^{-1}h\to 0$, we can assert (c.f. Collomb and Härdle(1986), Silverman(1979):

$$\text{(A.1a)} \qquad \sup |\hat{f}_h(x) - f(x)|\; I[f(x)>b] = O_p[(N^{1-(\varepsilon/2)}h^k)^{-1/2}]$$

$$\text{(A.1b)} \qquad \sup |\hat{f}_h'(x) - f'(x)|\; I[f(x)>b] = O_p[(N^{1-(\varepsilon/2)}h^{k+2})^{-1/2}]$$

for any $\varepsilon>0$.

Now define two (unobservable) "estimators" which are related to $\hat{\delta}$. First, define the estimator $\bar{\delta}$ based on trimming with respect to the true density value:

$$\text{(A.2)} \qquad \bar{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_h(x_i)y_i I_i$$

where $I_i \equiv I[f(x_i)>b]$, $i=1,\ldots,N$. Next, define a linearization $\tilde{\delta}$:

$$\text{(A.3)} \qquad \tilde{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\lambda}_h(x_i)y_i I_i$$

where

$$\text{(A.4)} \qquad \hat{\lambda}_h = \ell - \frac{\hat{f}_h'}{f} - \frac{\hat{f}_h}{f}\ell$$

Proof of Theorem 3.1: The proof consists of four distinct steps, summarized as:

Step 1. Linearization: $\sqrt{N}(\tilde{\delta} - \bar{\delta}) = o_p(1)$.

Step 2. Asymptotic Normality: $\sqrt{N}[\tilde{\delta} - E(\tilde{\delta})]$ has a limiting normal distribution with mean 0 and variance Σ.

Step 3. Bias: $[E(\tilde{\delta}) - \delta] = o(N^{-1/2})$.

Step 4. Trimming: $\sqrt{N}(\hat{\delta} - \delta)$ has the same limiting distribution as $\sqrt{N}(\bar{\delta} - \delta)$.

The combination of Steps 1-4 yields Theorem 3.1

Step 1. Linearization: Some arithmetic gives

$$\sqrt{N}(\tilde{\delta} - \bar{\delta}) = N^{-1/2} \sum_i \frac{[f(x_i) - \hat{f}_h(x_i)][\hat{f}_h'(x_i) - f'(x_i)]}{\hat{f}_h(x_i)\ f(x_i)} y_i I_i$$

$$- N^{-1/2} \sum_i \frac{[f(x_i) - \hat{f}_h(x_i)]^2}{\hat{f}_h(x_i)\ f(x_i)} \ell(x_i) y_i I_i$$

so that by (A.1a), there is a constant c_f such that with high probability

$$\sqrt{N}|\tilde{\delta} - \bar{\delta}| \le \frac{\sqrt{N}}{b^2 - bc_f(N^{1-(\varepsilon/2)}h^k)^{-1/2}} \sup_x[|f-\hat{f}_h|I] \sup_x[|\hat{f}_h' - f'|I] \frac{\Sigma|y_i|I_i}{N}$$

$$+ \frac{\sqrt{N}}{b^2 - bc_f(N^{1-(\varepsilon/2)}h^k)^{-1/2}} \sup_x[|f-\hat{f}_h|I]^2 \frac{\Sigma|\ell(x_i)y_i|I_i}{N}$$

The terms $N^{-1}\Sigma|y_i|I_i$ and $N^{-1}\Sigma|\ell(x_i)y_i|I_i$ are bounded in probability by Chebyshev's inequality. Consequently, from (A.1a,b) we have that

$$\sqrt{N}|\tilde{\delta} - \bar{\delta}| = O_p(b^{-2}N^{-(1/2)+(\varepsilon/2)}h^{-(2k+2)/2}) = o_p(1)$$

since $b^2N^{1-(\varepsilon/2)}h^k \to \infty$ and $b^4N^{1-\varepsilon}h^{2k+2} \to \infty$ by condition (ii).

Step 2: Asymptotic Normality: Write the linearization $\tilde{\delta}$ as

$$\tilde{\delta} = \tilde{\delta}_0 + \tilde{\delta}_1 + \tilde{\delta}_2$$

where

$$\tilde{\delta}_0 = N^{-1} \sum_i \ell(x_i) y_i I_i$$

$$\tilde{\delta}_1 = - N^{-1} \sum_i \frac{\hat{f}_h{}'(x_i)}{f(x_i)} y_i I_i$$

$$\tilde{\delta}_2 = - N^{-1} \sum_i \frac{\hat{f}_h(x_i)}{f(x_i)} \ell(x_i) y_i I_i$$

We show that $\sqrt{N}[\tilde{\delta} - E(\tilde{\delta})]$ has a limiting normal distribution, by showing that $\tilde{\delta}_0$, $\tilde{\delta}_1$ and $\tilde{\delta}_2$ are $\sqrt{N}$ equivalent to (ordinary) sample averages, and appealing to standard central limit theory. Throughout this section, we denote $(y_i, x_i) = v_i$. For $\tilde{\delta}_0$, we have that

$$\text{(A.5)} \qquad \sqrt{N}[\tilde{\delta}_0 - E(\tilde{\delta}_0)] = N^{-1/2}\left[\sum_{i=1}^{N} \{r_0(v_i) - E[r_0(v)]\}\right] + o_p(1)$$

where $r_0(v) = \ell(x)y$,

since $\mathrm{Var}(\ell y)$ exists and $b \to 0$ and $N \to \infty$.

To analyze $\tilde{\delta}_1$ and $\tilde{\delta}_2$, we approximate them by U-statistics. The U-statistic related to $\tilde{\delta}_1$ can be written as

$$U_1 = \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{1N}(v_i, v_j)$$

with

$$p_{1N}(v_i, v_j) = - \frac{1}{2} \left[\frac{1}{h}\right]^{k+1} K'\left[\frac{x_i - x_j}{h}\right] \left[\frac{y_i I_i}{f(x_i)} - \frac{y_j I_j}{f(x_j)}\right]$$

where $K' \equiv \partial K/\partial u$. Note that by symmetry of $K(.)$, we have

$$\sqrt{N}[\tilde{\delta}_1 - E(\tilde{\delta}_1)] = \sqrt{N}[U_1 - E(U_1)] - N^{-1}\{\sqrt{N}[U_1 - E(U_1)]\}$$

The second term in this expansion will converge in probability to zero provided $\sqrt{N}[U_1 - E(U_1)]$ has a limiting distribution, which we show later. Consequently, we have

$$(A.6) \qquad \sqrt{N}[\tilde{\delta}_1 - E(\tilde{\delta}_1)] = \sqrt{N}[U_1 - E(U_1)] + o_p(1)$$

The U-statistic related to $\tilde{\delta}_2$ is

$$U_2 = \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{2N}(v_i, v_j)$$

where

$$p_{2N}(v_i, v_j) = -\frac{1}{2} \left[\frac{1}{h}\right]^k K\left[\frac{x_i - x_j}{h}\right] \left[\frac{\ell(x_i)y_iI_i}{f(x_i)} + \frac{\ell(x_j)\, y_jI_j}{f(x_j)}\right]$$

U_2 is related to $\tilde{\delta}_2$ via

$$\sqrt{N}[\tilde{\delta}_2 - E(\tilde{\delta}_2)] = \sqrt{N}[U_2 - E(U_2)] - N^{-1}\{\sqrt{N}[U_2 - E(U_2)]\}$$

$$+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\frac{1}{Nh^k}\right] K(0) \left[\left[\frac{\ell(x_i)y_iI_i}{f(x_i)}\right] - E\left[\frac{\ell(x_i)y_iI_i}{f(x_i)}\right]\right]$$

As above, the second term of the expansion will converge in probability to zero provided $\sqrt{N}[U_2 - E(U_2)]$ has a limiting distribution, as shown later. The third term converges in probability to zero, because its variance is bounded by $K(0)^2(1/Nh^k)^2(h/b)^2 E[(\ell(x_i)y_i)^2 I_i) = o(1)$, since $Nh^k \to \infty$ and $h/b \to 0$. Therefore, we have

$$(A.7) \qquad \sqrt{N}[\tilde{\delta}_2 - E(\tilde{\delta}_2)] = \sqrt{N}[U_2 - E(U_2)] + o_p(1)$$

The analysis of U_1 and U_2 is quite similar, so we only present the details for U_1. We note that U_1 is a U-statistic with varying kernel (for instance, see Nolan and Pollard (1987)), since p_{1N} depends on N through the bandwidth h. Asymptotic normality of U_1 can be shown using Lemma 3.1 of Powell, Stock and Stoker(1987), which states that if $E[|p_{1N}(v_i, v_j)|^2] = o(N)$, then

$$(A.8) \qquad \sqrt{N}[U_1 - E(U_1)] = N^{-1/2}\left[\sum_{i=1}^{N} \{r_{1N}(v_i) - E[r_{1N}(v)]\}\right] + o_p(1)$$

$$\text{where } r_{1N}(v) = 2E[p_{1N}(v, v_j)|v]$$

This condition is implied by (ii): let $M_1(x_i) \equiv E(y_iI_i|x_i)$ and

$M_2(x_i) \equiv E(y_i^2 I_i | x_i)$, then

$$E[|p_{1N}(v_i,v_j)|^2]$$

$$\leq \frac{1}{4b^2h^{2k+2}} \int \left| K'\left[\frac{x_i - x_j}{h}\right]\right|^2 [M_2(x_i)+M_2(x_j)-2M_1(x_i)M_1(x_j)]\ f(x_i)f(x_j)dx_i dx_j$$

$$= \frac{1}{4b^2h^{k+2}} \int | K'(u)|^2 [M_2(x_j)+M_2(x_i+hu)-2M_1(x_i)M_1(x_i+hu)]\ f(x_i)f(x_i+hu)dx_i du$$

$$= O(b^{-2}h^{-(k+2)}) = O[N(b^2Nh^{k+2})^{-1}] = o(N)$$

since $b^2Nh^{k+2} \to \infty$ is implied by condition (ii). Therefore, (A.8) is valid.

Now let $b^* = \sup_{x,u}\{f(x+hu) | f(x)=b \text{ and } |u| \leq 1\}$ and $I_i^* = I[f(x_i)>b^*]$. Note that by construction, if u is such that $|u| \leq 1$, then $I[f(x_i+hu)>b]-I_i^* \neq 0$ only when $(1-I_i^*)=1$, and that $b^* \to 0$, $h/b^* \to 0$ as $b \to 0$ and $h \to 0$. Now write $r_{1N}(v_i) = E[2p_{1N}(v_i,v_j)|v_i]$ as:

$$r_{1N}(v_i) = -\frac{1}{h^{k+1}} \int K'\left[\frac{x_i - x}{h}\right] \left[\frac{y_i I_i}{f(x_i)} - \frac{m(x)I[f(x)>b]}{f(x)}\right] f(x)dx$$

$$= \frac{y_i I_i}{f(x_i)} \int \frac{1}{h} K'(u)\ f(x_i+hu)du$$

$$- I_i^* \int \frac{1}{h} K'(u)\ m(x_i+hu)du$$

$$- (1 - I_i^*) \int \frac{1}{h} K'(u)\ m(x_i+hu)\{I[f(x_i+hu)>b]-I_i^*\}du$$

$$= - \frac{y_i I_i}{f(x_i)} \int K(u)\ f'(x_i+hu)du$$

$$+ I_i^* \int K(u)\ m'(x_i+hu)du \quad + (1 - I_i^*)\ a(x_i;h,b)$$

where

$$a(x_i;h,b) = -\int \frac{1}{h} K'(u)\ m(x_i+hu)\{I[f(x_i+hu)>b]-I_i^*\}du$$

Now, if $r_1(v_i)$ is defined as

$$r_1(v_i) = \ell(x_i)y_i + m'(x_i)$$

then the difference between $r_{1N}(v_i)$ and $r_1(v_i)$ is

$$t_{1N}(v_i) = r_{1N}(v_i) - r_1(v_i) = -\frac{y_i I_i}{f(x_i)} \int K(u)[f'(x_i+hu) - f'(x_i)]du$$

$$+ I_i^* \int K(u)[m'(x_i+hu) - m'(x_i)]du \; + (1 - I_i)\ell(x_i)y_i$$

$$+ (1 - I_i^*)m'(x_i) + (1 - I_i^*)\, a(x_i;h,b)$$

The second moment $E[t_{1N}(v_i)^2]$ vanishes as $N\to\infty$: By the Lipschitz conditions of Assumption 7 the second moment of $[y_i I_i/f(x_i)]\int K(u)[f'(x_i+hu)-f'(x_i)]du$, is bounded by $(h/b)^2(\int|u|K(u)du)^2 E[y^2\omega_{f'}(x)^2]=O[(h/b)^2]=o(1)$. The second moment of $I_i^*\int K(u)[m'(x_i+hu)-m'(x_i)]du$ is bounded by $h^2(\int|u|K(u)du)^2 E[\omega_{m'}(x)^2]=O(h^2)=o(1)$. The second moments of $(1-I_i)\ell(x_i)y_i$ and $(1-I_i^*)m'(x_i)$ vanish by Assumption 4 since $b\to 0$ and $b^*\to 0$. Finally, the second moment of $(1-I_i^*)a(x_i;h,b)$ vanishes if the second moment of $a(x;h,b)$ exists, since $b^*\to 0$. Now consider the ℓ^{th} component $a_\ell(x;h,b)$ of $a(x;h,b)$, and define the "marginal kernel" $K_{(\ell)}=\int K(u)du_\ell$ and the "conditional kernel" $K_\ell=K/K_{(\ell)}$. For a given x, integrating $a_\ell(x;h,b)$ by parts absorbs h^{-1}, and shows that $a_\ell(x;h,b)$ is the sum of two terms: first the expectation (w.r.t. $K(u)$) of $m_\ell'(x+hu)\{I[f(x+hu)>b]-I[f(x)>b^*]\}$, and second the expectation (w.r.t. $K_{(\ell)}$) of $K_\ell m(x+hu)$ over u values such that $f(x+hu)=b$. Because the variances of m' and y exist, the second moment (over x) of each of these terms exists, so that $E[a_\ell(x;h,b)^2]$ exists. Consequently, since the second moment of each component of $a(x;h,b)$ exists, we have that the second moment of $(1-I_i^*)a(x;h,b)$ vanishes, so that $E[t_{1N}(v_i)^2]=o(1)$.

This fact is sufficient to show asymptotic normality of U_1, because

$$\text{(A.9)} \qquad N^{-1/2}\sum_{i=1}^{N}\{r_{1N}(v_i)-E[r_{1N}(v_i)]\} = N^{-1/2}\sum_{i=1}^{N}\{r_1(v_i)-E[r_1(v_i)]\} + N^{-1/2}\sum_{i=1}^{N}\{t_{1N}(v_i)-E[t_{1N}(v_i)]\}\ ,$$

and the last term converges in probability to zero, since its variance is bounded by $E[t_{1N}(v)^2] = o(1)$. Consequently, combining (A.9), (A.8) and (A.6), we conclude that

(A.10) $$\sqrt{N}[\tilde{\delta}_1 - E(\tilde{\delta}_1)] = N^{-1/2}\left[\sum_{i=1}^{N}\{r_1(v_i) - E[r_1(v)]\}\right] + o_p(1)$$

where $r_1(v) = \ell(x)y + m'(x)$

The U-statistic representation U_2 of $\tilde{\delta}_2$ is analyzed in a similar fashion. In particular, $E[|p_{2N}(v_i,v_j)|^2]=o(N)$ if $b^2Nh^k\to\infty$, which is implied by condition (ii). By an analogous argument, $[U_2 - E(U_2)]$ is shown to be $\sqrt{N}$ equivalent to the average of $[r_2(v_i) - E(r_2(v))]$, where $r_2(v)=-[\ell(x)y + \ell(x)m(x)]$. Combining this with (A.6) permits us to conclude that

(A.11) $$\sqrt{N}[\tilde{\delta}_2 - E(\tilde{\delta}_2)] = N^{-1/2}\left[\sum_{i=1}^{N}\{r_2(v_i) - E[r_2(v)]\}\right] + o_p(1)$$

where $r_2(v) = -[\ell(x)y + \ell(x)m(x)]$

Combining (A.5), (A.10) and (A.11) then yields Step 2, as

$$\sqrt{N}[\tilde{\delta} - E(\tilde{\delta})] = N^{-1/2}\left[\sum_{i=1}^{N}\{r(v_i) - E[r(v)]\}\right] + o_p(1) \quad ,$$

with $r(v) \equiv r_0(v) + r_1(v) + r_2(v) = m'(x) + [y - m(x)]\ell(x)$,

with $r(v) \equiv r(y,x)$ in the statement of Theorem 3.1.

Step 3: Bias: Expand the bias of $\tilde{\delta}$ as

$$E(\tilde{\delta}) - \delta = \tau_{1N} - \tau_{2N} + \tau_{3N}$$

where

$$\tau_{1N} = E\left[N^{-1}\sum_{i=1}^{N} \ell(x_i)y_iI_i\right] - \delta$$

$$\tau_{2N} = E\left[N^{-1}\sum_{i=1}^{N} [\hat{f}_h'(x_i) - f'(x_i)]\frac{y_iI_i}{f(x_i)}\right]$$

$$\tau_{3N} = E\left[N^{-1}\sum_{i=1}^{N} [\hat{f}_h(x_i) - f(x_i)]\frac{\ell(x_i)y_iI_i}{f(x_i)}\right]$$

Let A_N, B_N be defined as before. Then

$$\tau_{1N} = \int_{A_N} \ell(x)m(x)f(x)dx - \int \ell(x)m(x)f(x)dx = \int_{B_N} m(x)f'(x)dx = o(N^{-1/2})$$

by Assumption 8.

We only show that $\tau_{2N}=o(N^{-1/2})$, with the proof of $\tau_{3N}=o(N^{-1/2})$ quite similar. Let ι denote an index set $(\ell_1,\ldots,\ell_k)$, where $\Sigma\ell_j=p$. For a k-vector $u=(u_1,\ldots,u_k)$, define $u^\iota=u_1^{\ell_1}u_2^{\ell_2}\ldots u_k^{\ell k}$, and let $f_\iota^{(p)}$ denote the p^{th} partial derivative of f with respect to the u components indicated by ι, namely $f_\iota^{(p)}=\partial^p f/(\partial u)^\iota$. By partial integration we have

$$\tau_{2N} = \int_{A_N} m(x) \int K(u)\ [f'(x-hu) - f'(x)]\ du\ dx$$

$$= \int_{A_N} m(x) \sum_\iota \int K(u)\ h^{p-1} f_\iota^{(p)}(\xi)u^\iota\ du\ dx$$

where the summation is over all index sets ι with $\Sigma\ell_j=p$, and where ξ lies on the line segment between x and x-hu. Thus

$$\tau_{2N} = h^{p-1}\int_{A_N} m(x) \sum_\iota f_\iota^{(p)}(x)\int K(u)u^\iota\ du\ dx$$

$$+ h^{p-1}\int_{A_N} m(x) \sum_\iota \int K(u)\ [f_\iota^{(p)}(\xi)-f_\iota^{(p)}(x)]\ u^\iota\ du\ dx \ = O(h^{p-1})$$

by Assumption 9. Therefore, by condition (iii), we have $\tau_{2N}=O[N^{-1/2}(N^{1/2}h^{p-1})]=o(N^{-1/2})$, as required. Thus $E(\tilde{\delta})-\delta=o(N^{-1/2})$.

Step 4: <u>Trimming</u>: Steps 1-3 show that $\sqrt{N}(\bar{\delta} - \delta)$ has a limiting normal distribution with mean 0 and variance-covariance matrix Σ. We now show the same result for $\hat{\delta}$. Let $\Phi(Z)$ denote the cummulative distribution function of the normal distribution with mean 0 and variance Σ.

Let $c_N=c_f(N^{1-(\varepsilon/2)}h^k)^{-1/2}$, where c_f is an upper bound consistent with (A.1a). Define two new trimming bounds as $b_u=b+c_N$ and $b_\ell=b-c_N$, and the associated trimmed kernel estimators:

$$\bar{\delta}_u = \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_h(x_i) y_i 1[f(x_i) > b_u]$$

$$\bar{\delta}_\ell = \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_h(x_i) y_i 1[f(x_i) > b_\ell]$$

Since $b^{-1} c_N \to 0$ by condition (ii), $\bar{\delta}_u$ and $\bar{\delta}_\ell$ each obey the tenets of steps 1-3, so $\sqrt{N}(\bar{\delta}_u - \delta)$ and $\sqrt{N}(\bar{\delta}_\ell - \delta)$ each have a limiting normal distribution with mean 0 and variance Σ (whose cumulative distribution function is denoted Φ). Moreover, by construction, we have that

$$\mathrm{Prob}\{\sqrt{N}(\bar{\delta}_u - \delta) \le Z,\ |\hat{f}_h(x_i) - f(x_i)| \le c_N,\ i=1,\ldots,N\}$$

$$\le \mathrm{Prob}\{\sqrt{N}(\hat{\delta} - \delta) \le Z,\ |\hat{f}_h(x_i) - f(x_i)| \le c_N,\ i=1,\ldots,N\}$$

$$\le \mathrm{Prob}\{\sqrt{N}(\bar{\delta}_\ell - \delta) \le Z,\ |\hat{f}_h(x_i) - f(x_i)| \le c_N,\ i=1,\ldots,N\}$$

By (A.1a), as $N \to \infty$, $\mathrm{Prob}\{\sup|\hat{f}_h(x_i) - f(x_i)| > c_N\} \to 0$. Consequently, as $N \to \infty$, we have

$$\Phi(Z) \le \lim \mathrm{Prob}\{\sqrt{N}(\hat{\delta} - \delta) \le Z\} \le \Phi(Z)$$

so that $\lim \mathrm{Prob}\{\sqrt{N}(\hat{\delta} - \delta) \le Z\} = \Phi(Z)$.

<u>Proof of Theorem 3.2</u>: The estimator $\hat{r}_{hi}$ is constructed by direct estimation of the U-statistic component structure of $\hat{\delta}$. First, set $\hat{r}_{0N}(v_i) = \hat{\ell}_{hi}(x_i) y_i \hat{I}_i$. Next define $\hat{p}_{1N}(v_i, v_j)$ and $\hat{p}_{2N}(v_i, v_j)$ by replacing $f(x_i)$, $\ell(x_i)$, I_i, $f(x_j)$, $\ell(x_j)$, I_j by $\hat{f}_{hi}(x_i)$, $\hat{\ell}_{hi}(x_i)$, $\hat{I}_i$, $\hat{f}_{hj}(x_j)$, $\hat{\ell}_{hj}(x_j)$, $\hat{I}_j$ respectively, in the formulae defining $p_{1N}(v_i, v_j)$ and $p_{2N}(v_i, v_j)$. Finally, define $\hat{r}_{1N}(v_i) = 2N^{-1} \Sigma_j p_{1N}(v_i, v_j)$ and $\hat{r}_{2N}(v_i) = 2N^{-1} \Sigma_j p_{2N}(v_i, v_j)$. By techniques similar to Collomb and Härdle(1986) and Silverman(1979), by construction we have that $\sup|[\hat{r}_{0N}(v_i) + \hat{r}_{1N}(v_i) + \hat{r}_{2N}(v_i) - r(v_i)] I_i| = o_p(1)$. The estimator $\hat{r}_{hi}$ of (3.5) is just $\hat{r}_{hi} \equiv \hat{r}_{0N}(v_i) + \hat{r}_{1N}(v_i) + \hat{r}_{2N}(v_i)$.

By an argument similar to that of Step 4 above, it suffices to prove consistency of $\bar{\Sigma} = N^{-1} \Sigma\, \hat{r}_{hi} \hat{r}_{hi}^T 1_i - \hat{\delta}\hat{\delta}^T$. Set $r_i \equiv r(y_i, x_i)$, then

$$\bar{\Sigma} - [E(rr^T) - \delta\delta^T]$$

$$= N^{-1}\Sigma(\hat{r}_{hi}-r_i)(\hat{r}_{hi}-r_i)^T I_i + N^{-1}\Sigma\, r_i(\hat{r}_{hi}-r_i)^T I_i + N^{-1}\Sigma(\hat{r}_{hi}-r_i)r_i^T$$

$$- N^{-1}\Sigma r_i r_i^T(1-I_i) + N^{-1}\Sigma r_i r_i^T - E(rr^T) - \hat{\delta}\hat{\delta}^T + \delta\delta^T$$

$$= o_p(1)$$

since $\sup|\hat{r}_{hi}-r_i|I_i=o_p(1)$, $E(rr^T)$ exists, $\text{Prob}\{f(x)\le b\}=o(1)$ and $\hat{\delta}$ is consistent for δ.

Proof of Theorem 3.3:

With $z_j=x_j'\delta$, define $d_j=\hat{z}_j-z_j$. Since $f_1(z)\ge b_1>0$, $d_j=x_j^T(\hat{\delta}-\delta)$ $=x_j^T O_p(N^{-1/2})$, so that $\sup_j\{d_j\}=O_p(N^{-(1/2)})$. Define the kernel regression function estimator using z_j instead of $\hat{z}_i$ as

$$\tilde{g}_{h'}(z) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{h'}\right] K_1\left[\frac{z - z_j}{h'}\right] y_j \Big/ \tilde{f}_{1h'}(z)$$

where $\tilde{f}_{1h'}$ is the density estimator:

$$\tilde{f}_{1h'}(z) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{h'}\right] K_1\left[\frac{z - z_j}{h'}\right]$$

When $h'\sim N^{-1/5}$, it is a standard result (e.g. Schuster 1972) that $N^{2/5}[\tilde{g}_{h'}(z)-g(z)]$ has the asymptotic distribution given in Theorem 3.3. Consequently, the result follows if $\hat{g}_{h'}(z)-\tilde{g}_{h'}(z)=o_p(N^{-2/5})$.

First consider $\hat{f}_{1h'}-\tilde{f}_{1h'}$. By applying the triangle inequality to the Taylor expansion of $\hat{f}_{1h'}(z)$ we have

$$|\hat{f}_{1h'}(z) - \tilde{f}_{1h'}(z)| \le |\sup\{d_j\}|\ |\tilde{f}'_{h'}(z)|$$

$$+ \sup\{d_j^2\}\ h'^{-1}|N^{-1}h'^{-3}\sum_j (K_1)''[(z-\xi_j)/h']|$$

where ξ_j lies between $\hat{z}_j$ and z_j. Therefore $\hat{f}_{1h'}(z)-\tilde{f}_{1h'}(z)=O_p(N^{-(1/2)})$ $=o_p(N^{-2/5})$. By a similar argument, $\hat{f}_{1h'}(z)\hat{g}_{h'}(z)-\tilde{f}_{1h'}(z)\tilde{g}_{h'}(z)=o_p(N^{-2/5})$. Consequently $\hat{g}_{h'}(z)-\tilde{g}_{h'}(z)=o_p(N^{-2/5})$. QED Theorem 3.3.

References

Beran, R.(1977), "Adaptive Estimates for Autoregressive Processes," Annals of the Institute of Statistical Mathematics, 28, 77-89.

Bickel, P.(1982), "On Adaptive Estimation," Annals of Statistics, 10, 647-671.

Box, G.E.P. and D.R. Cox(1964), "An Analysis of Transformations," Journal of the Royal Statistical Society, Ser. B, 26, 211-252.

Breiman, L. and J.H. Friedman(1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," Journal of the American Statistical Association, 80, 580-619.

Bronstein, I.N. and K.A. Semandjajew(1974), Taschenbuch der Mathematik, B.G. Teubner, Verlagsgesellschaft.

Collomb, G. and W. Härdle(1986), "Strong Uniform Convergence Rates in Robust Nonparametric Time Series Analysis and Prediction: Kernel Regression Estimation from Dependent Observations," Stochastic Processes and Their Applications, 23, 77-89.

Friedman, J.H. and W. Stuetzle(1981), "Projection Pursuit Regression," Journal of the American Statistical Association, 76, 817-823.

Gasser, T., H.G. Müller and V. Mammitzsch(1985), "Kernels for Nonparametric Curve Estimation," Journal of the Royal Statistical Society, Ser. B, 47.

Härdle, W.(1987), "XploRe - A Computing Environment for Exploratory Regression," in Statistical Data Analysis Based on the L1 Norm, edited by Y. Dodge, North Holland.

Härdle, W. and J. S. Marron(1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," Annals of Statistics, 13, 1465-1481.

Härdle W. and J.S. Marron(1987), "Semiparametric Comparison of Regression Curves," Working Paper, Sonderforschungsbereich 303, Universitat Bonn, February.

Hastie, T. and R. Tibshirani(1986), "Generalized Additive Models (with discussion)," Statistical Science, 1, 297-318.

Ichimura, H. (1987), Estimation of Single Index Models, doctoral dissertation, Massachusetts Institute of Technology.

Manski, C.F.(1984), "Adaptive Estimation of Nonlinear Regression Models," draft, Department of Economics, University of Wisconsin, Madison.

Manski, C.F. and D. McFadden(1981), Structural Analysis of Discrete Data with Econometric Applications, ed., Cambridge, Massachusetts, MIT Press.

McCullagh, P. and J.A. Nelder(1983), Generalized Linear Models, Chapman and Hall, London.

Nolan D. and D. Pollard(1987), "U-Processes: Rates of Convergence," Annals of Statistics, 15, 780-799.

Powell, J.L.(1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," Econometrica, 54, 1435-1460.

Powell, J.L., J.H. Stock and T.M. Stoker(1987) "Semiparametric Estimation of Index Coefficients," MIT School of Management Working Paper WP #1973-86, revised October.

Robinson, P.M.(1987) "Root N-Consistent Semiparametric Regression," London School of Economics Econometrics Projects, Discussion Paper R.9, forthcoming Econometrica.

Schuster, E.F.(1972), "Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Number of Distinct Points," Annals of Mathematical Statistics, 43, 84-88.

Silverman, B.W.(1979), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and Its Derivatives," Annals of Statistics, 6, 177-184 (Addendum 1980, Annals of Statistics, 8, 1175-1176).

Stoker, T.M.(1986), "Consistent Estimation of Scaled Coefficients," Econometrica, 54, 1461-1481.

Stone, C.J.(1980), "Optimal Rates of Convergence for Nonparametric Estimators," Annals of Statistics, 8, 1348-1360.

Stone, C.J.(1986),"The Dimensionality Reduction Principle for Generalized Additive Models," Annals of Statistics, 14, 590-606.

Table 1: ADE Estimation of the Linear Model (5.1)
$N = 50 \quad \alpha = 1\%$

	$h = 1.5$	Low Bandwidth $h = 1.0$	High Bandwidth $h = 2.0$
$\hat{\delta}_1$	.9713 (.2775)	.5059 (.1556)	1.0403 (.2634)
$\hat{\delta}_2$	2.0389 (.2476)	1.0587 (.1542)	2.0410 (.3408)
$\hat{\delta}_3$	3.0366 (.2940)	1.6328 (.2163)	3.1858 (.3779)
$\hat{\delta}_4$	4.0158 (.2661)	2.1354 (.2568)	4.1292 (.4385)
b	.0015	.0077	.0005

Table 2: ADE Estimation of the Sine Model (5.2)
N = 100 α = 5%

	h = .9	Low Bandwidth h = .7	High Bandwidth h = 1.5	Known Density
$\hat{\delta}_1$	.1134 (.0960)	.0428 (.0772)	.1921 (.1350)	.1329 (.1228)
$\hat{\delta}_2$	.1356 (.1093)	.0449 (.0640)	.1982 (.1283)	.1340 (.1192)
$\hat{\delta}_3$	.1154 (.1008)	.0529 (.0841)	.1837 (.1169)	.1330 (.1145)
$\hat{\delta}_4$	.1303 (.0972)	.0591 (.0957)	.2042 (.1098)	.1324 (.1251)
b	.0117	.0321	.0017	

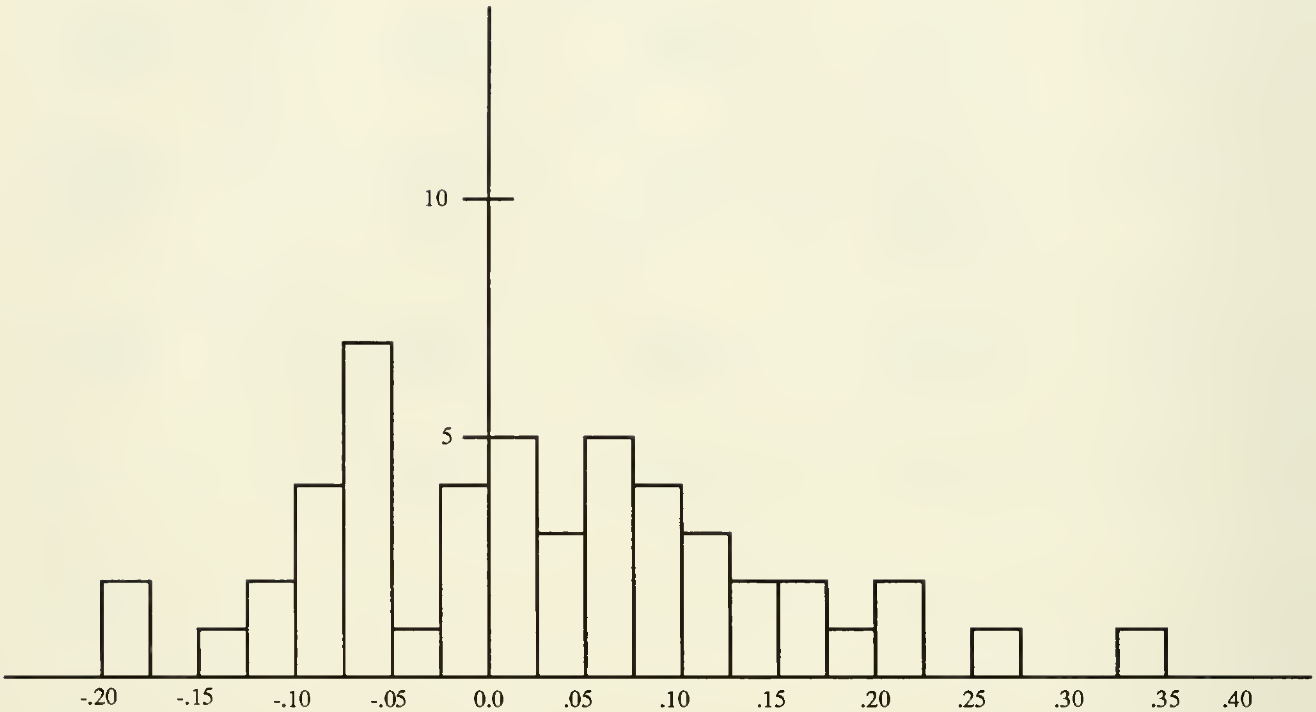

Figure 1: Histogram of Observed Differences Between Average Squared Errors: ADE Regression and Multivariate Nonparametric Regression.

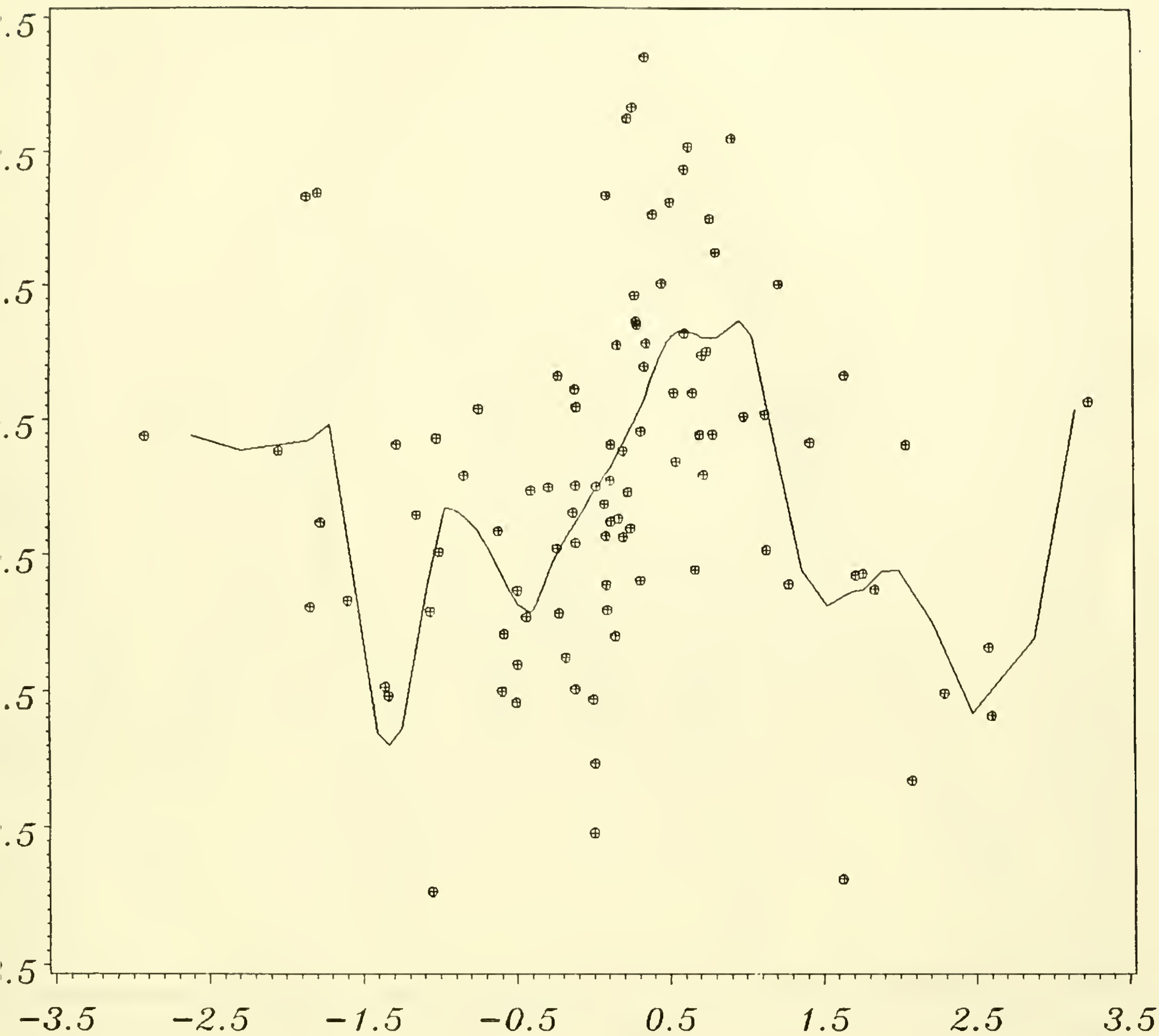

Figure 2: Basic Data and ADE Regression

———— : $\hat{g}_{h'}$

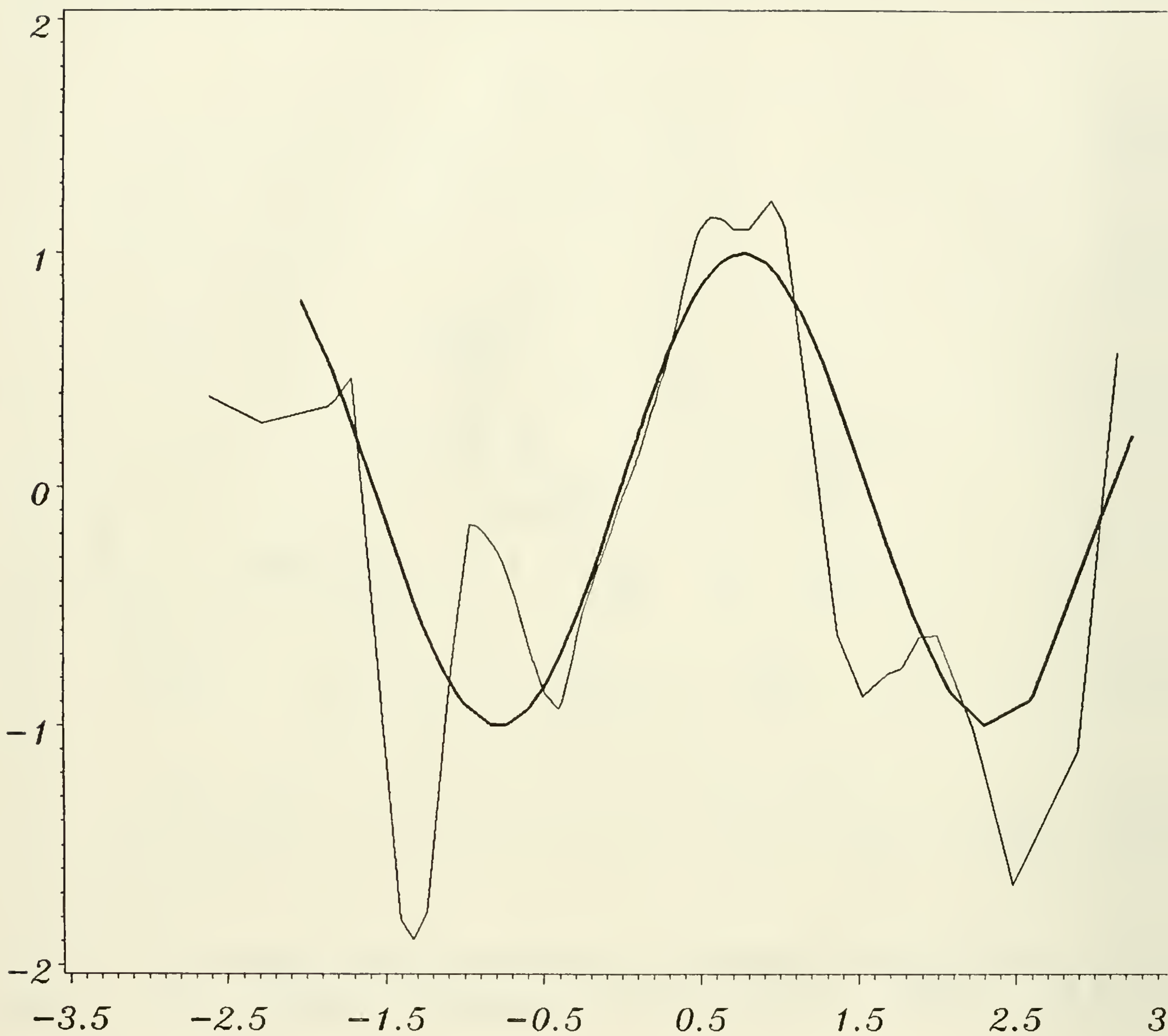

Figure 3: ADE Regression and the True Regression Curve

———————— : $\hat{g}_{h'}$

———————— : m

Figure 4: ADE Regression and Regression Based on Known δ

——— : $\hat{g}_{h'}$

━━━ : $\tilde{g}_{h'}$

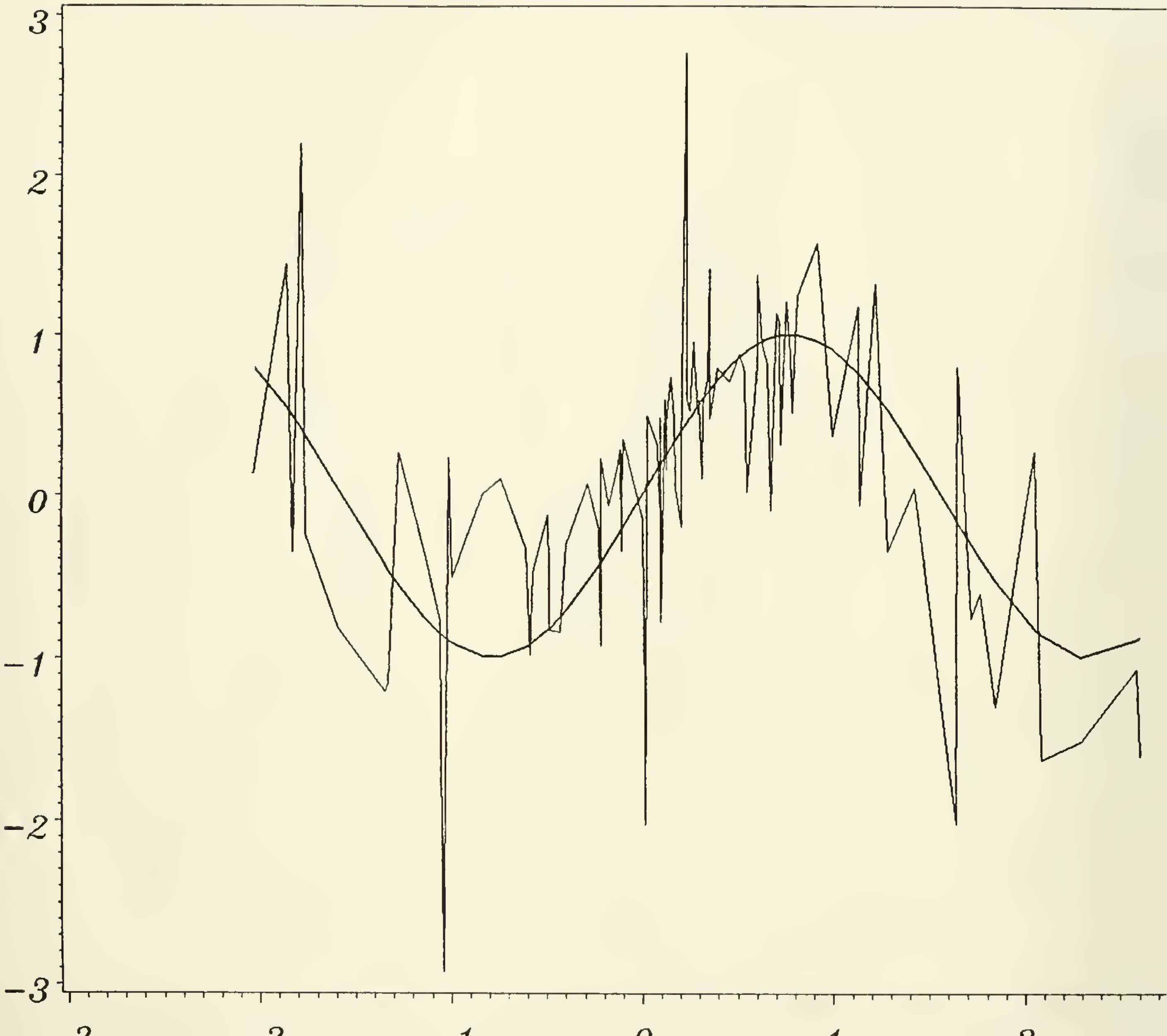

Figure 5: Multivariate Nonparametric Regression and the True Regression Curve

———— : $\tilde{m}_{h''}$

━━━━ : m

4/16/88

Date Due

MAY 3 1		
OCT 2 9 1997		